# Distributed robust statistical learning: A Byzantine mirror descent algorithm

**Mihailo Jovanović**
ee.usc.edu/mihailo

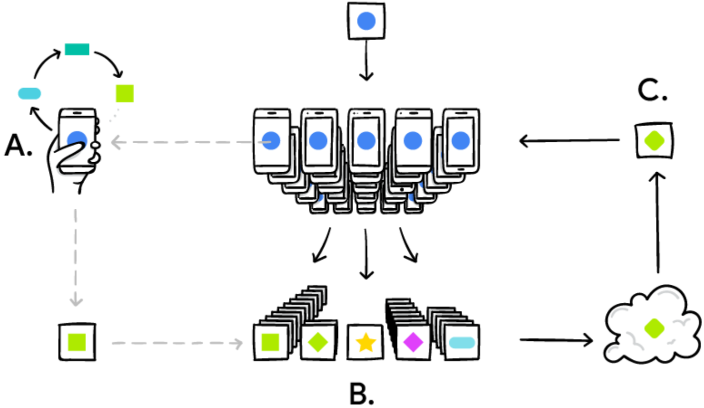
work of

Dongsheng Ding and Xiaohan Wei

**58th IEEE Conference on Decision and Control**

# Motivating application

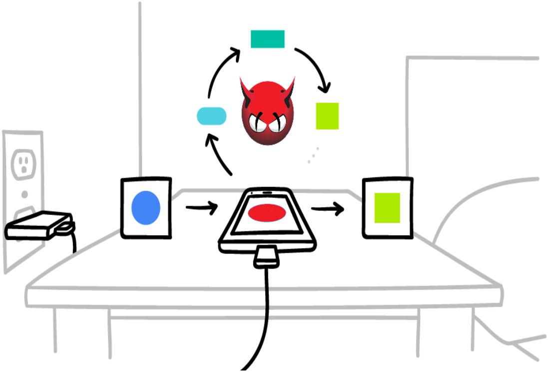
- FEDERATED LEARNING




A. Worker machine    B. Master machine    C. Shared model

Google AI, Blog ’17

# Byzantine fault



- FAULT SOURCES
  - Machine failures
  - Communication errors
  - Malicious users

# Byzantine failure model

- STOCHASTIC LEARNING PROBLEM

$$
\begin{aligned}
\underset{w}{\text{minimize}} \quad & F(w) \; := \; \mathbb{E}_{z \sim \mathcal{D}}\left(f(w;z)\right) \\
\text{subject to} \quad & w \; \in \; \mathcal{W} \; \subset \; \mathbb{R}^d
\end{aligned}
$$

# Byzantine failure model

- STOCHASTIC LEARNING PROBLEM

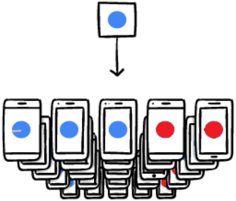
$$\begin{aligned} \underset{w}{\text{minimize}} \quad & F(w) \; := \; \mathbb{E}_{z \sim \mathcal{D}}\left(f(w; z)\right) \\ \text{subject to} \quad & w \; \in \; \mathcal{W} \; \subset \; \mathbb{R}^d \end{aligned}$$

$1$ master and $m$ workers

- $\star$ $z_t^i \sim \mathcal{D}$, $i \in \{1, \ldots, m\}$

$m$ gradients at time $t$

$$\nabla_t^i \; := \; \begin{cases} \nabla f(w_t; z_t^i) & \text{normal machine} \\ \text{arbitrary} & \text{Byzantine machine} \end{cases}$$



# Identification of "good" workers

- MEDIAN AGGREGATION
  - robust to outliers

| sequence | median |
|---|---|
| $1,\ 3,\ 3,\ 6,\ 7,\ 8,\ 10$ | $6$ |
| $10^{-10},\ 3,\ 3,\ 6,\ 7,\ 8,\ 10^{10}$ | $6$ |

# Euclidean setting

- BYZANTINE SGD

$$w_{t+1} \;:=\; \underset{w \in \mathcal{W}}{\operatorname{argmin}} \; F(w_t) \;+\; \eta \, \langle \xi_t, w - w_t \rangle \;+\; \boxed{\frac{1}{2} \|w - w_t\|_2^2}$$

$\xi_t$ – stochastic estimate of the gradient $\nabla F(w_t)$

$$\xi_t \;=\; \frac{1}{m} \sum_{i \in \Omega_t} \nabla_t^i$$

$\Omega_t$ – set of "good" workers

# Convergence

- CONVEX SMOOTH OBJECTIVE FUNCTION

$$F(\bar{w}) - F(w^\star) \le \tilde{O}\left(C\left(\frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right)\right) \text{ w.h.p.}$$

- $\star$ $\bar{w} := \frac{1}{T}\sum_{t=1}^{T} w_{t+1}$
- $\star$ $T$ – total number of iterations
- $\star$ $\|w - w'\|_2 \le W$ for all $w, w' \in \mathcal{W} \subset \mathbb{R}^d$
- $\star$ $\|\nabla_t^i - \nabla_t\|_2 \le C$ – gradient norm bound for "good" workers

Alistarh, Allen-Zhu, Li, NeurIPS '18

# Example

- LINEAR REGRESSION

$$f(w; z) = \frac{1}{2}(x^T w - y)^2, \quad w \in \mathcal{W} \subset \mathbb{R}^d$$

⋆ $z := (x, y)$ – data generated by $y = x^T w^\star + \xi$

$$x(i) \sim \{-1, 1\}, \quad \xi \sim \mathcal{N}(0, \sigma^2)$$

**dimension-dependent gradient norm bound**

$$\mathbb{E}\left(\|\nabla_t^i - \nabla_t\|_2\right) \leq \sqrt{(d-1)\, W^2 + d\,\sigma^2}$$

Yin, Chen, Kannan, Bartlett, ICML ’18

# Exploiting problem geometry to improve the dimension dependence

# Bregman divergence

$$D(x, y) \ := \ \Phi(x) \ - \ \Phi(y) \ - \ \nabla\Phi(y)^T(x \ - \ y)$$

⋆ $\Phi$ – differentiable, $1$-strongly convex w.r.t. $\|\cdot\|$

# Bregman divergence

$$D(x, y) \;:=\; \Phi(x) \;-\; \Phi(y) \;-\; \nabla\Phi(y)^T(x \;-\; y)$$

⋆ $\Phi$ – differentiable, $1$-strongly convex w.r.t. $\|\cdot\|$

- EXAMPLES

  ⋆ $\Phi(x) \;=\; \frac{1}{2}\,\|x\|_2^2$ strongly convex w.r.t. $\|\cdot\|_2$

$$D(x, y) \;=\; \frac{1}{2}\,\|x \;-\; y\|_2^2$$

# Bregman divergence

$$D(x,y) := \Phi(x) - \Phi(y) - \nabla\Phi(y)^T(x - y)$$

⋆ $\Phi$ – differentiable, $1$-strongly convex w.r.t. $\|\cdot\|$

- EXAMPLES

⋆ $\Phi(x) = \frac{1}{2}\|x\|_2^2$ strongly convex w.r.t. $\|\cdot\|_2$

$$D(x,y) = \frac{1}{2}\|x - y\|_2^2$$

⋆ $\Phi(x) = \sum_i x(i)\log x(i)$ strongly convex w.r.t. $\|\cdot\|_1$

$$D(x,y) = \sum_i x(i)\log\frac{x(i)}{y(i)}$$ KL divergence

Bubeck, Found. Trends Mach. Learn. '15

# Non-Euclidean setting

- BYZANTINE MIRROR DESCENT

$$w_{t+1} := \operatorname*{argmin}_{w \in \mathcal{W}} F(w_t) + \eta \langle \xi_t, w - w_t \rangle + \boxed{D(w, w_t)}$$

$\xi_t$ – stochastic estimate of the gradient $\nabla F(w_t)$

$$\xi_t = \frac{1}{m} \sum_{i \in \Omega_t} \nabla_t^i$$

$\Omega_t$ – set of "good" workers

- IMPORTANT QUANTITIES

  - $\nabla_t^1, \dots, \nabla_t^m$ – gradients (normal or Byzantine)

  - $A_t^1, \dots, A_t^m$ – gradient related values

$$A_t^i := \sum_{k=1}^{t} \langle \nabla_k^i, w_k - w_1 \rangle$$

  - $B_t^1, \dots, B_t^m$ – accumulated gradients

$$B_t^i := \sum_{k=1}^{t} \nabla_k^i$$

used to update the set of "good" workers

# Convergence result

- CONVEX AND $L$-SMOOTH OBJECTIVE FUNCTION

$$F(\bar{w}) - F(w^\star) \leq \frac{2R^2}{\eta T} + \frac{8\sqrt{2}WC\Delta(1+4\alpha\sqrt{m})}{\sqrt{mT}} + \eta\left(\frac{32C^2\Delta^2}{m} + 64\alpha^2C^2\right) \quad \text{w.h.p.}$$

$$\sup_{w \in \mathcal{W}} D(w, w_1) \leq R^2$$

$\star$

$$\Delta = \Theta\left(\sqrt{\log \frac{mT}{\delta}}\right)$$

$$\eta \leq \frac{1}{2L}$$

matches standard mirror descent for $C = 0$

# Optimal rate

- OPTIMAL STEPSIZE

$$\eta = \begin{cases} \min\left(\frac{1}{\alpha C\sqrt{T}}, \frac{1}{2L}\right), & \alpha \geq \frac{1}{\sqrt{m}} \\ \min\left(\frac{1}{C}\sqrt{\frac{m}{T}}, \frac{1}{2L}\right), & \alpha < \frac{1}{\sqrt{m}} \end{cases}$$

$$F(\bar{w}) - F(w^\star) \leq \tilde{O}\left(C\left(\frac{R^2}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right)\right) \quad \text{w.h.p.}$$

matches the rate of $\begin{cases} \text{Byzantine SGD}, & \alpha \neq 0 \\ \text{batch SGD}, & \alpha = 0 \end{cases}$

# Probability simplex

$$\mathcal{W} \;:=\; \left\{ w \in \mathbb{R}^d,\; \|w\|_1 = 1,\; w \geq 0 \right\}$$

$$\|\cdot\| \;=\; \|\cdot\|_1, \quad \|\cdot\|_* \;=\; \|\cdot\|_\infty$$

- KL DIVERGENCE

$$D(x, y) \;=\; \sum_i x(i) \log \frac{x(i)}{y(i)}$$

  - $\star$ $\|\nabla_t^i - \nabla_t\|_\infty \leq C$ – dimension-independent bound
  - $\star$ $w_1 = (\frac{1}{d}, \cdots, \frac{1}{d})$ – uniform initialization

$$D(w^\star, w_1) \;\leq\; \log d \;=\; R^2$$

- CONVEX AND $L$-SMOOTH OBJECTIVE FUNCTION

$$F(\bar{w}) - F(w^\star) \leq \frac{2\log d}{\eta T} + \frac{8C\Delta(\sqrt{mT} + 4\alpha m\sqrt{T})}{mT} + \eta\left(\frac{4C^2\Delta^2}{m} + 32\alpha^2 C^2\right) \quad \text{w.h.p.}$$

$C$ – dimension-independent constant

OPTIMAL STEPSIZE $\eta$

$$F(\bar{w}) - F(w^*) \leq \tilde{O}\left(\frac{\log d}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right) \quad \text{w.h.p.}$$

# Summary

- RESULTS
  - ⋆ Byzantine mirror descent
  - ⋆ Probability simplex: nearly dimension-free

- ONGOING EFFORT
  - ⋆ Problems with constraints
  - ⋆ Byzantine primal-dual algorithm

# Extra slides

# Concentration bounds

- **Gradient bias**

$$
\begin{aligned}
|E_1| &= \left| \sum_{t=1}^{T} \sum_{i \in \Omega_t} \langle \nabla_t^i - \nabla_t, w_t - w^\star \rangle \right| \\
&\leq 4WC\Delta\sqrt{2Tm} + 16\alpha m W C \Delta \sqrt{2T} \quad \text{w.h.p.}
\end{aligned}
$$

- **Gradient variance**

$$
\begin{aligned}
E_2 &= \frac{1}{T} \sum_{t=1}^{T} \left\| \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t) \right\|_*^2 \\
&\leq \frac{16 C^2 \Delta^2}{m} + 32 \alpha^2 C^2 \quad \text{w.h.p.}
\end{aligned}
$$

  - ⋆ $I_A = 4WC\Delta\sqrt{2T}$
  - ⋆ $I_B = 4C\Delta\sqrt{2T}$
  - ⋆ $\Delta = R + 2\sqrt{2 \log \frac{8\sqrt{2}mT}{\delta}}$

Rakhlin, Sridharan, COLT '17

# Convergence analysis

- **Convex and $L$-smooth objective**

$$\frac{1}{mT}\sum_{t=1}^{T}\sum_{i\in\Omega_t}(F(w_{t+1})-F(w^\star)) \leq \eta E_2 + \frac{R^2}{\eta T} - \frac{E_1}{mT}$$

$$\leq \underbrace{\eta C^2\left(\frac{16\Delta^2}{m}+32\alpha^2\right)}_{\text{variance}} + \underbrace{\frac{R^2}{\eta T}}_{\text{error}} + \underbrace{C\frac{4\sqrt{2}W\Delta}{\sqrt{mT}} + \alpha C\frac{16\sqrt{2}W\Delta}{\sqrt{T}}}_{\text{bias}}$$

  - ⋆ $\|w-w'\|\leq W$ for all $w, w'\in\mathcal{W}$
  - ⋆ $D(w^\star, w_1)\leq R^2$
  - ⋆ $\|\nabla_t^i-\nabla_t\|_*\leq C$
  - ⋆ $\eta\leq\frac{1}{2L}$